eip ESCUELA INTERNACIONAL POSGRADOS

Máster Avanzado de Programación en Python para Hacking, BigData y Machine Learning

Programación Python para BigData

LECCIÓN 08

Apache Spark con PySpark [2/2]

ÍNDICE

- ✓ Introducción
- ✓ Objetivos
- ✓ Spark MLib
- ✓ Spark GraphX
- ✓ Diferencia entre Hadoop y Spark
- ✓ Conclusiones

INTRODUCCIÓN

En esta lección aprenderemos a trabajar con Spark usando una librería de Python como es PySpark, para ello usaremos una imagen de Docker usada en la lección anterior y veremos los distintos usos con los que podemos trabajar.

OBJETIVOS

Al finalizar esta lección serás capaz de:

1. Conocer el uso de Spark MLib
2. Conocer el uso de Spark GraphX
3. Breve introducción a Hadoop

Spark MLib

Permite realizar Machine Learning usando spark.

- **Algoritmo DecisionTreeClassifier**

```
In [29]: from pyspark.ml.classification import DecisionTreeClassifier

         dt = DecisionTreeClassifier(labelCol='Survived',
                                     featuresCol='features',
                                     maxDepth=5)
```

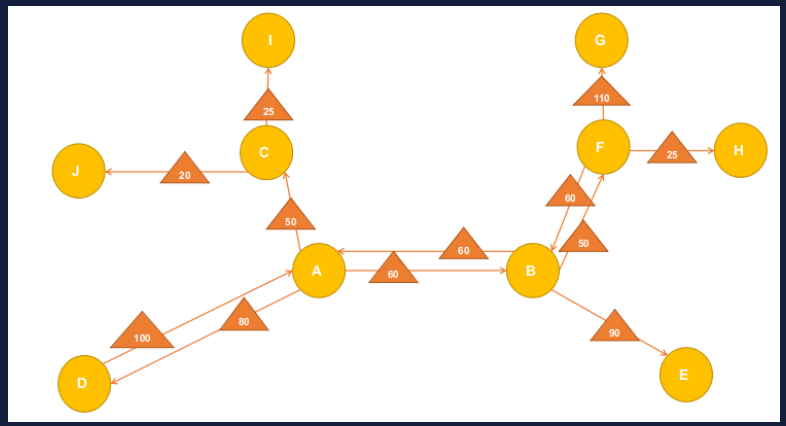
- **Algoritmo Gradient-boosted tree classifier**

```
In [35]: from pyspark.ml.classification import GBTClassifier

         # Entrenar el modelo GBT
         gbt = GBTClassifier(labelCol="Survived", featuresCol="features", maxIter=10)
```

Spark GraphX

- API de Python, Java y Scala: GraphFrames proporciona interfaces API comunes para los tres lenguajes. Es la primera vez que todos los algoritmos implementados en GraphX se pueden usar en Python y Java.
- Consultas potentes: GraphFrames permite consultas breves, al igual que las consultas potentes en Spark SQL y DataFrame.
- Guardar y cargar modelos de gráficos: GraphFrames es totalmente compatible con las fuentes de datos de estructura DataFrame, lo que permite el uso de Parquet, JSON y CSV familiares para leer y escribir gráficos.



Diferencia entre Hadoop y Spark

1. Propósito: Hadoop distribuye grandes conjuntos de datos a múltiples nodos en un clúster compuesto por varias computadoras para su almacenamiento. Spark es una herramienta especialmente utilizada para procesar macrodatos en almacenamiento distribuido.

2. Distinta implementación de los dos

3. Velocidad de procesamiento de datos: Spark tiene las ventajas de Hadoop y MapReduce que son más adecuadas para la minería de datos y el aprendizaje automático que requieren iteración.

4. Recuperación de la seguridad de los datos.

CONCLUSIONES

1. Spark MLib sirve para realizar aprendizaje automático.

2. Spark GraphX sirve para el procesamiento general de gráficos, basándose en la teoría de grafos.

3. La principal ventaja que presenta Spark frente a Hadoop es que usa la memoria para el procesamiento lo que permite una mayor rapidez de los procesamientos.

MUCHAS GRACIAS POR SU ATENCIÓN

eip ESCUELA INTERNACIONAL POSGRADOS

imaniega@grupomainjobs.com

Isabel Maniega
https://www.linkedin.com/in/isabel-maniega-cuadrado-40a8356b/

twitter.com/eiposgrados

facebook.com/eiposgrados

instagram.com/eiposgrados